Gregor K Schroeder

Data Scientist | Statistician
UC San Diego *MS Statistics*, UC Santa Barbara *BS Statistics*
GREGORKSCHROEDER@GMAIL.COM | LINKEDIN.COM/IN/GREGOR-SCHROEDER | (619) 871-1011

TECHNICAL SKILLS AND TECHNOLOGIES

Azure ML, GCP, Git, Google BigQuery, Microsoft Azure, PowerBI, Python, R, Streamlit, SQL, VBA for Excel
ArcGIS, Docker, MLFlow, MLlib, Observable JS, PySpark, scikit-learn, Shiny, TensorFlow, Quarto, XGBoost

---

WORK EXPERIENCE

SANDAG | MANAGER OF REGIONAL MODELS Oct 2023 - Present

• Manger of the San Diego Association of Governments (SANDAG) Estimates & Forecasts team. Currently supervising a technical team of six responsible for all population, housing, and employment products at the agency. Implemented GitHub repository standards and workflows including issue-based development practices, pull request review workflows, and Wiki-based publicly available documentation for internal, private, and public stakeholders.
• Leading three software re-architecture projects transitioning SANDAG's short-range population and housing estimates model, regional long-range population and housing forecasting model, and subregional long-range population and housing allocation model into a fully Python based platform. These projects represent the team's main products requiring understanding of years of loosely documented business rules and prioritization of technical debt that has accrued while still maintaining consistent product release schedules.
• Evaluating historical forecast errors for products released by SANDAG since 1970 enabling the development of empirical prediction intervals to better inform users of SANDAG's long-range population and housing forecasts. Researching, testing, and developing model-based prediction intervals for long-range population and housing forecast models to supplement empirical prediction intervals.

BIOLOGICAL DYNAMICS | SENIOR DATA SCIENTIST Oct 2021 – July 2023

• Designed and implemented a production machine learning platform using Python on the Microsoft Azure ML Platform, incorporating integrated project management tools, Azure DevOps, Git, data versioning, Docker containers, scalable on-demand cloud computing, ML model traceability, and production deployment of machine learning solutions at scale.
• Developed a sophisticated machine learning pipeline that allowed for global search across combinations of feature engineering strategies and algorithms, including regularized regression, support vector machines (SVM), tree-based methods, boosting methods, and deep learning neural networks, using both nested and non-nested cross-validation techniques to emphasize performance in independent validation.
• Presented the organization's machine learning strategy and results to journal editors, C-level executives, Board of Directors, and outside investor groups as the lead Data Scientist, owning the company's algorithm development process. This led to the successful publication of research in Nature Communications Medicine and the development of commercial tests for early cancer detection.

SANDAG | PRINCIPAL RESEARCHER+MODELER Sep 2012 - Sep 2021

• Spearheaded database, data warehouse, and business intelligence architecture and development as the lead developer. Designed, created, and maintained large-scale production SQL databases and ETL (Extract, Transform, Load) methods for SANDAG's micro-simulated transportation model, which served as the primary source for all transportation-modeling datasets used in the agency's 2021 Regional Plan.

• Implemented a complex and scalable multi-dimensional Iterative Proportional Fitting (IPF) algorithm to weight transportation survey data by eligible persons attributes at varying geographic resolutions, feeding into nested binary/multiclass choice models.

• Supported and assisted in survey design, data collection, weighting, and estimation of nested binary and multiclass classification used to construct SANDAG’s micro-simulated activity-based transportation model.

CONECTRIC NETWORKS | DATA SCIENCE CONSULTANT Apr 2018 - Feb 2019

• Led the development of Conectric Network's automated Google Cloud Platform soft audit tool as the lead Data Scientist and Python developer.

• Automated the extraction, aggregation, interpolation, and statistical decomposition of utility energy meter kilowatt-hour (kWh) readings for client hotel sites into base kWh, weather-dependent controllable kWh, and weather-independent controllable kWh.

• Developed an automated time-series model toolkit with empirical prediction intervals to forecast utility meter energy usage. The results were subsequently used by Conectric to determine potential profitability and further engagement with client sites. Leveraged expertise in statistical analysis, time-series modeling, and Python programming to deliver a scalable and reliable solution for the client.

PUBLICATIONS

Hinestrosa, J.P., Sears, R.C., Dhani, H., Lewis, J. M., Schroeder, G. et al. Development of a blood-based extracellular vesicle classifier for detection of early-stage pancreatic ductal adenocarcinoma. Commun Med 3, 146 (2023). https://doi.org/10.1038/s43856-023-00351-4

Hinestrosa, J.P., Kurzrock, R., Lewis, J.M., Schork, N. J., Schroeder, G. et al. Early-stage multi-cancer detection using an extracellular vesicle protein-based blood test. Commun Med 2, 29 (2022). https://doi.org/10.1038/s43856-022-00088-6